論文 / 著書情報
Article / Book Information

題目(和文)	タスクに応じた単語分割
Title(English)	Task-Oriented Word Segmentation
著者(和文)	平岡達也
Author(English)	Tatsuya Hiraoka
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第11829号, 授与年月日:2022年3月26日, 学位の種別:課程博士, 審査員:岡崎 直観,徳永 健伸,篠田 浩一,宮崎 純,井上 中順
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第11829号, Conferred date:2022/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	審査の要旨
Type(English)	Exam Summary

（博士課程）

論文審査の要旨及び審査員

報告番号	甲第　　号		学位申請者氏名	平岡 達也	
論文審査 審査員		氏　名	職　名		氏　名	職　名
	主査	岡崎 直観	教授	審査員	井上 中順	准教授
	審査員	德永 健伸	教授			
		篠田 浩一	教授			
		宮崎 純	教授			

論文審査の要旨（2000字程度）

本論文は、「Task-Oriented Word Segmentation」と題し、英文7章から構成されている。自然言語処理では、システムに入力されたテキストを単語列に分解（以降「単語分割」と呼ぶ）し、文書分類や感情極性分析、機械翻訳などのタスク（以降「後段タスク」と呼ぶ）を解くのが通常である。ところが、単語分割の処理と、後段のタスクを解く処理は独立に設計されるため、後段タスクにとって最適な単語分割が得られる保証はない。本論文では、後段モデルと単語分割モデルを同時に最適化することで、後段タスクのデータセットとモデルに応じて、適切な単語分割モデルを学習する手法を提案している。

第1章「Introduction」では、本研究の背景、目的、貢献を述べている。まず、自然言語処理における単語分割処理と、その問題点を指摘している。その後、本論文で提案する手法の概要およびその貢献を説明している。

第2章「Related Work and Preliminary」では、本研究で用いられる自然言語処理技術の従来研究の概要および本研究の位置づけが述べられている。教師無し単語分割を概観したのち、Byte Pair Encoding (BPE)、言語モデルによる単語分割手法（SentencePiece）を詳説し、これらの単語分割の違いについて、例を用いながら説明している。複数の単語分割結果を利用する手法として、BPE-dropoutとサブワード正則化を説明している。最後に、後段タスクとして、BERTやLSTMなどの深層ニューラルネットワークを用いた文書分類や機械翻訳の概要を説明している。

第3章「OpTok: Optimizing Tokenization for Text Classification」では、文書分類のために単語分割を最適化する手法（OpTok）を提案している。OpTokは文ベクトルに単語分割の確率をかけ合わせることで、単語分割モデルを後段モデルの学習の中に組み込む。後段タスクの学習における損失値に基づき、単語分割モデルが適切な単語分割に高い確率を与えるように言語モデルのパラメータを更新する。この章では、モデルの訓練を安定化させるための工夫も述べられている。ただし、OpTokは入力文を一つの文ベクトルにエンコードすることを仮定しているため、そのままでは機械翻訳などの後段タスクに適用できない。

第4章「OpTok4AT: Optimizing Tokenization for Various Tasks」では、OpTokを様々な後段タスクに適用するために拡張した手法として、OpTok4ATを提案している。OpTok4ATでは、後段モデルの学習で計算される損失に対して単語分割の確率をかけ合わせるため、OpTokのように文ベクトルを利用する必要がない。これにより、OpTok4ATは文ベクトルを用いない後段タスクにも適用できる。この章ではさらに、提案手法をサブワード正則化と組み合わせることで、メモリ使用量を抑えた学習方法を説明している。また、機械翻訳タスクのように、複数の入力文が与えられるタスクにOpTok4ATを適用する場合の学習方法にも言及している。

第5章「Experiments」では、文書分類タスクと機械翻訳タスクにおける複数言語での実験とその結果を述べている。実験の結果より、OpTokは文書分類タスク、OpTok4ATは文書分類タスクと機械翻訳タスクで、単語分割の最適化を行わない場合と比べて高い性能を示すことを報告している。

第6章「Discussion」では、複数の実験設定を用いてOpTokとOpTok4ATの振る舞いを分析している。具体的には、提案手法で単語分割のみを更新した場合であっても、後段タスクの性能向上が得られることを報告している。また、提案手法によって獲得された単語分割を分析し、提案手法がタスクに応じて異なる単語分割を獲得していることを報告している。続いて、ある後段タスクで学習した単語分割を、異なる後段タスクに転用する実験から、提案手法が後段タスクに特化した単語分割を獲得できることを示している。提案手法をマルチタスク学習に適用した実験では、提案手法が各タスクの特徴を考慮した単語分割を獲得し、性能向上に寄与したことを報告している。さらに、提案手法をBERTに適用した場合にも、性能向上が得られることを示している。また、提案手法を調整するためのハイパ

ーパラメータの影響について分析している。

第7章「Conclusion」では、本論文の提案手法（OpTok、OpTok4AT）の設計と得られた実験結果について概観し、タスクに応じた単語分割の最適化が自然言語処理における性能向上に有効であると結論づけている。また、提案手法は学習に時間がかかるという問題点を提起し、より効率の良い学習方法を用いた高速化を将来展望として挙げている。

本論文では、後段タスクのデータセットとモデルに応じて適切な単語分割モデルを学習する手法を提案し、3言語の文書分類タスクと、7言語対の機械翻訳タスクの実験により、提案手法が適切な単語分割を学習することで、後段タスクの性能の向上に寄与することを実証した。本論文は、単語の境界をどのように捉えるべきか、コンピュータの情報処理の視点から検討し、工学の発展に寄与した。一方で、本研究の知見を踏まえて、人間が単語の境界をどのように考えるべきか、すなわち形態論への展開も期待できる。以上要するに、本論文は博士（工学）の学位論文として十分価値あるものと認める。